# Population genomics reveals evolution and variation of *Saccharomyces cerevisiae* in the human and insects gut

Matteo Ramazzotti[1]*, Irene Stefanini[2]*, Monica Di Paola[3]*, Carlotta De Filippo[4], Lisa Rizzetto[5], Luisa Berná[6], Leonardo Dapporto[3], Damariz Rivero[3], Noemi Tocci[5], Tobias Weil[5], Marcello S. Lenucci[7], Paolo Lionetti[8], Duccio Cavalieri[3§]

**Affiliations:**

[1]*Department of Experimental and Clinical Biomedical Sciences, University of Florence, Florence, Italy.*

[2]*Division of Biomedical Sciences, University of Warwick, Coventry, United Kingdom*

[3]*Department of Biology, University of Florence, Florence, Italy*

[4]*Institute of Agricultural Biology and Biotechnology, National Research Council (CNR), Pisa, Italy.*

[5]*Fondazione E. Mach, Research and Innovation Centre, San Michele all' Adige (Trento), Italy.*

[6]*Unidad de Biología Molecular, Institut Pasteur de Montevideo, Montevideo, Uruguay*

[7]*Dipartimento di Scienze e Tecnologie Biologiche ed Ambientali (Di.S.Te.B.A.), Università del Salento, Lecce, Italy*

[8]*Department of Neuroscience, Psychology, Drug Research and Child Health, Meyer Children Hospital, University of Florence, Florence, Italy.*

**These authors contributed equally to the work.*

[§]*Corresponding author: Duccio Cavalieri, via della Madonna del Piano 6, 50019 Sesto Fiorentino, Florence, Italy; telephone number: 0039 055 4574717; e-mail: duccio.cavalieri@unifi.it.*

# Originality-Significance Statement

Here we report the first detailed genomic and phenotypic characterization of *Saccharomyces cerevisiae* strains isolated from human intestine. This comprehensive analysis details the possible features of gut strains enabling survival and persistence in this complex and clinically relevant environment.

Microsatellite-based typing of strains revealed the existence of three main clusters and a clear evidence of clonal expansion or colonization within the human gut. Whole-genome sequencing combined with a plethora of genome-wide analyses, such as phylogenomics on functionally conserved protein coding sequences, ancestry analysis, copy number variations, introgression of non-reference genes and degree of polymorphisms were used. Such analyses confirmed the phylogenetic relationships among strains isolated either from human or wasp guts (used for control) and provided functional evidence of genetic differences between strains isolated from gut and non-gut environments. Strains were further characterized at phenotypic level to investigate canonical features as well as cell-wall composition and host immune system recognition pattern.

Overall, our analyses allowed us to draw relevant insights on how yeast can persist in the human gut, finding that cell wall composition, sporulation efficiency and nutrient competition could be primarily involved. Interestingly, sporulation efficiency and strain ancestry were linked with strain immunogenicity.

This study provides evidence for selection of specific traits and gene networks in strains inhabiting the gut, allowing us to improve our understanding on how this commensal organism establishes symbiotic interaction with potential benefit to the host, suggesting a potential role of humans in the evolution of this yeast.

# Summary

The quest to discover the variety of ecological niches inhabited by *Saccharomyces cerevisiae* has led to research in areas as diverse as wineries, oak trees, and insect guts. The discovery of fungal

communities in the human gastrointestinal tract suggested the host's gut as a potential reservoir for yeast adaptation. Here we report the existence of yeast populations associated with the human gut (HG) that differ from those isolated from other human body sites. Phylogenetic analysis on 12 microsatellite loci and 1,715 combined CDSs from whole-genome sequencing revealed three subclusters of HG strains with further evidence of clonal colonization within the host's gut. The presence of such subclusters was supported by other genomic features, such as copy number variation, absence/introgressions of CDSs and relative polymorphism frequency. Functional analysis of CDSs specific of the different subclusters suggested possible alterations in cell wall composition and sporulation features. The phenotypic analysis combined with immunological profiling of these strains further showed that sporulation was related with strain-specific genomic characteristics in the immune recognition pattern. We conclude that both genetic and environmental factors involved in cell wall remodeling and sporulation are the main drivers of adaptation in *S. cerevisiae* populations in the human gut.

## Running title:

Human gut as *Saccharomyces cerevisiae* niche.

# Introduction

Despite in-depth knowledge of the genetic, molecular, and phenotypic traits regulating the physiology of *Saccharomyces cerevisiae*, the forces shaping its origin and evolution are still debated. The long lasting association of *S. cerevisiae* with human activities (3,150 B.C) (Cavalieri et al., 2003; Fay and Benavides, 2005; Legras et al., 2007; Goddard et al., 2010; Liti, 2015) has led to the idea that its wide use in fermentation has caused its domestication. Few studies have examined the population biology of *S. cerevisiae*, its genetic variation and evolution (Fay and Benavides, 2005; Legras et al., 2007; Aa et al., 2006; Ezov et al., 2006; Ruderfer et al., 2006; Zhu et al., 2016; Almeida et al., 2017; Zhu et al., 2017). However, whole genome analysis classified *S. cerevisiae* strains accordingly to the isolation source (Oak, Wine/European, or West African) and the type of human activity from which they derived (Dunn et al., 2012), suggesting that the population structure of this yeast consists of a few isolated lineages and mosaicism (Liti et al., 2009; Liti 2015). Despite the available reports on the worldwide distribution and genetic diversity of *S. cerevisiae* (Peter et al., 2018), it is still unclear how and to what extent human intervention has shaped the baker's yeast population structure (Liti et al., 2009; Schacherer et al., 2009; Goddard et al., 2010; Borneman et al., 2011). Recent discovery of the role of wasps' gut as a winter reservoir and potential evolutionary niche for *S. cerevisiae* indicated how the ecological cycle of this microorganism is just beginning to be understood (Stefanini et al., 2012). The possible evolution of *S. cerevisiae* in the gastrointestinal tract of the animals is not limited to wasps or fruit flies (Knight et al., 2015), but probably extends to vertebrates and finally to mammals. Since yeasts are ubiquitous in human-related environments, especially in association with food production (leavening bread and fermenting wine and beer), it is not surprising that yeasts, including *S. cerevisiae*, populate the skin (Findley et al., 2013), mucosal membranes (Sobel et al., 1993; Liguori et al., 2016), breast milk (Boix-Amorós et al., 2017), blood (Smith et al., 2002; Swinne et al., 2009), and respiratory (Tawfik et al., 1989; Williams et al., 2007) and intestinal tracts (Ott et al., 2008; Liguori et al., 2016; Sokol et al., 2017) of humans (Rizzetto et al., 2014). Exposure to fungi is constant, and recent studies have highlighted that the *mycobiota* is a significant player in host-

microbe interactions, being also able to shape both innate and adaptive immunity (Hardison and Brown, 2012; Sancho et al., 2012; Nguyen et al., 2015; Romani et al., 2015; Strati et al., 2016a). *Candida albicans* is the best-known human commensal, but *S. cerevisiae* has also been shown to be a potential colonizer of the human intestinal tract (Angebault et a., 2013; Xu et al., 1999; Strati et al., 2016a). A recent hypothesis is that human environment-associated *S. cerevisiae* strains gave rise to clinical *S. cerevisiae* causing colonization/infection (Strope et al., 2015). Noticeably, this species was identified in both immunocompromised patients (Muñoz et al., 2005.; Williams et al., 2007; Swinne et al., 2009) and healthy subjects (Sobel et al., 1993; Smith et al., 2002).

Currently, the fungal community is emerging as a player of the symbiotic interactions in the gut ecosystem (Ott et al., 2008; Liguori et al., 2016; Sokol et al., 2017). Studies investigating fungal communities in chronic inflammation, especially in Inflammatory Bowel Diseases (IBD), described a more heterogeneous mycobiota in patients affected by Crohn's Disease (CD) compared to healthy subjects, suggesting a role for altered mycobiota as the causative agents of inflammatory diseases and "leaky gut" syndrome (Ott et al., 2008; Schulze and Sonnenborn, 2009; Li et al., 2013; Sokol et al., 2017). An association between gut inflammation and abundance of different yeast species, including *S. cerevisiae*, was found in mouse models of chemically-induced colitis (Iliev et al., 2012; Chiaro et al., 2017). Furthermore, Ott and collaborators showed an enrichment in *S. cerevisiae* strains in IBD faecal samples (Ott et al., 2008). A study on paediatric subjects reported a clear fungal dysbiosis in CD patients with an enrichment in *Candida spp.* and a reduction of *S. cerevisiae* in disease versus remission, thus proposing a positive role of *S. cerevisiae* colonization (Sokol et al., 2017). On the other hand, Liguori and colleagues (Liguori et al., 2016) observed that *S. cerevisiae* was enriched in the gut mucosa of CD patients, but this yeast species was also present in non-inflamed gut mucosa. Strikingly, one of the markers proposed to discriminate CD from Ulcerative Colitis (both IBDs) is CD positivity for anti-*Saccharomyces cerevisiae* antibodies (ASCA) (McKenzie et al., 1990; Quinton et al., 1998; Sendid 1998). All these studies have highlighted the relevance of *S. cerevisiae* in the intestinal homeostasis and inflammation, but the controversial results made difficult to have a definitive

answer on its role in the host health and disease. The wide genetic and phenotypic variability observed for this yeast and the different human-associated and environmental sources could be the causes behind the inability to draw general conclusions on the role of yeast in contributing to or protecting from disease (Rizzetto et al., 2016).

Thus, three general questions should be resolved: (i) are there differences in terms of abundance of strains between patients and healthy subjects?; (ii) are *S. cerevisiae* strains found in the human gut genotypically and phenotypically different from non-human environments?; (iii) how related are *S. cerevisiae* gut strains compared to those found on other human body sites and originated by environment or human activities?; (iv) are there genetic features characterizing gut strains?

In the current work, we investigated the genetic and phenotypic characteristics, variability, and population structure of *S. cerevisiae* isolates from the human gut of paediatric patients. In addition, to contextualize the implications of genetic variation on the ability of strains to survive and evolve in the harsh intestinal environment, we characterized the strains for traits potentially related to survival or adaptation to the gut environment and we investigated the impact of the observed genetic variations on the immunogenicity of these strains.

# Results

## Strain isolation, identification and typing

Aiming at the isolation of *S. cerevisiae* strains from human faeces, we plated faecal aliquots onto selective media (YPD supplemented with chloramphenicol, see Experimental Procedures) allowing the growth of only yeasts and fungi. To avoid mating of isolates after plating, faecal aliquots were diluted to have around 10 isolates per plate. Faecal samples were collected from 34 paediatric Crohn's Disease (CD) patients, 27 Ulcerative Colitis (UC) patients, 1 non-IBD paediatric patient, and 32 healthy children (HC). To identify the isolates' species, we analysed the ITS1-5.8S-ITS2 region using Sanger sequencing and assigned the species by comparing the results with all the ITS1-5.8S-ITS2 sequences available in the NCBI database. If the best match was a *S. cerevisiae* strain, the isolate was identified as *S. cerevisiae*. Using this procedure, we identified a total of 35

*S. cerevisiae* isolates, the great majority of which was found in CD patients, with 27 isolates from 6 different CD patients (17.7% of CD patients), 1 isolate from 1 UC (3.7%), 5 isolates from 1 non-IBD paediatric patient (nIBD), and 2 isolates from HC (3.1%) (see **Table 1** for isolate origin and nomenclature). However, we could not find any associations among the isolation of *S. cerevisiae* strains and the type of donor, neither comparing healthy donors against patients nor considering the different pathologies (Chi-squared test of independence p=0.523).

In order to investigate the origin of the isolated strains, we performed phylogenetic analysis based on 12 microsatellite loci (Legras et al., 2007) (**Figure 1a**). Neighbor-Joining clustering based on chord distances (Takezaki and Nei 1996) was used to assess the similarity among gut isolates and other strains from a wide variety of sources (soil, insect gut, grape, wine fermentation, bakery and clinical environment), for a total of 286 strains (Legras et al., 2007). The inclusion of strains isolated from wasp intestines was aimed at evaluating genetic similarities among strains isolated from the gastrointestinal tract.

Strain clustering was performed through K-means analysis on chord distances, with the best number of clusters (n=16) being identified as the smallest minimizing the distances among co-clustering strains (**Supplementary Figure 1**). Strains isolated from human faeces grouped in three clusters, with just three exceptions (the strains YB7, YB8, and YG12). A significant fraction of gut isolates (16 out of 35), grouped in a cluster (which we called the Human Gut 1, HG1 cluster) together with five laboratory strains S288C, FY10, W303, fl100, and fl200 (**Figure 1a,** beige shading). This cluster encompasses isolates (YB, YD and YE series) obtained from three CD patients (B, D and E respectively, **Table 1**). We identified two other sub-clusters of HG strains (indicated in blue in **Figure 1a**). One group (including 1 strain from an UC patient, 2 strains from CD patients, and 2 strains from HC), which we called HG2 (**Figure 1a**, pink shading) clustered closely to strains used in bread-making (Capa1, YS4, and MUCL42920, isolated from bread and beer). Another group of four HG strains (called HG3, blue shading in **Figure 1a**), isolated from a single CD patient (YH1-YH4), co-clustered with the SK1 laboratory strain. Interestingly, in two human subclusters we also found strains that had previously been isolated from the guts of wasp caught in the same geographical area where all patients and healthy subjects resided (Tuscany,

Italy; **Figure 1a**). An additional cluster included the strains isolated from the “B” patient, YB7 and YB8, that clustered with strains isolated from wine fermentation and grapes collected from the Tuscany (Italy) region where that patients resided (**Figure 1a**). Wasp intestines were previously shown to bear *S. cerevisiae* strains representing a large part of the local genetic diversity of this species (Stefanini et al., 2012), and the clustering based on microsatellite data confirms this observation, as wasp strains spread along the tree (**Figure 1a**). Some strains isolated from wasp intestines were similar to the gut isolates, probably because of the similarity of isolation sources (both are from intestines) and geographical location (faecal donors lived in the same region where the wasps were caught, Tuscany, Italy).

Noteworthy, the strains isolated from the same patient resulted more genetically similar than other strains belonging to the same cluster (i.e. among strains of the YB, YD, YE, and YH series), as indicated by their co-clustering (**Figure 1a**) and by the comparison of distances among strains isolated from the same faecal sample and co-clustering strains (**Figure 1b**). In fact, strains isolated from the same individual are more similar among each other than strains belonging to the same cluster (Mann-Whitney U test fdr<0.05, **Figure 1c**). To further support the observation that the distance among strains isolated from the same donor was lower than the distance among strains from other sources, we compared the genetic distances among strains isolated from the same donor and among strains randomly selected from the dataset of strains characterized by microsatellite analysis (10,000 sampling iterations for each group of strains isolated from the same donor). In more than 85% of the permutations, the strains isolated from the same donor were more similar among each other compared to re-sampled datasets (**Supplementary figure 2**). These results suggested either a clonal expansion of strains in the human intestine or multiple colonisations of a given faecal donor from multiple strains of the same source population.

## Strains sequencing, phylogenomics, and population analysis

To search for genetic traits characterizing HG strains, we obtained the entire genome sequence of a selection of 18 isolates from human gut, together with a set of 4 wasps gut strains, 5 strains from grapes, and 1 strain from wines collected in the Florence area, Italy; **Supplementary Table 1**).

Sequenced strains were selected to represent the broadest genetic variability observed in the microsatellite analysis (**Figure 1a**), by including one strain per major group and strains that did not cluster in any major group. As a quality control of our study, we sequenced a clone of S288c progenitor, the EM93 strain, isolated from decomposing fig (Mortimer and Johnston 1986). In order to disclose the genomic features shared by intestinal isolates, strains from wasp intestines were selected among those that were closer to human gut isolates according to the microsatellite data clustering (**Figure 1a**). In fact, by sampling a wider genetic variability of wasp strains, we would have also included strains simply vectored by wasps (Stefanini et al., 2012) and hence not showing features peculiar of intestinal isolates. For phylogenomic comparison purposes, we included in our analysis several strains isolated in previous studies from various geographical locations and isolation sources, also encompassing a large number of isolates from clinical specimens (Liti et al. 2009; Strope et al., 2015). Considering that such external reference strains were sequenced as haploids, to properly perform a phylogenetic comparison among strains, we sequenced meiotic segregants of candidate strains from this study that were able to sporulate and give viable spores. Other analyses (e.g. copy number variation, CDSs loss, new CDSs) were carried out on both meiotic segregants and parental (polyploid) strains.

The percentage identity of reads mapped on the S288c strain genome reference ranged between 93% and 97% (**Supplementary Table 2**). SNP calling at high coverage (lowest=46.7, strain YP4-40D; highest=304.7, strain YH1-28C) of newly sequenced strains with respect to the reference genome revealed a wide variability of SNPs/indels (Single Nucleotide Polymorphisms/insertions deletions), from a minimum of 538 total SNP/indels in the YE1 strain to a maximum of 101,330 in the YUC22 strain (**Supplementary Table 2**).

SNP calling using the S288c laboratory strain's genome as a reference was used to reconstruct coding sequences of strains and perform *per-CDS* reverse multiple sequence alignment in order to obtain a robust (see Experimental procedures) genome-wide alignment encompassing 1,715 CDSs from 144 strains, for a total of 245,990 informative loci (see Experimental procedures for further details). In order to fully exploit this information, we have created a publicly accessible database (http://bioserver2.sbsc.unifi.it/bioinfo/yeastpop/index.html) to inspect alignments and

produce phylogenetic trees based on concatenations of desired CDSs among the 1,715 available CDSs.

Phylogenetic analysis by approximately-maximum-likelihood tree construction confirmed the similarity among groups of strains isolated from human faeces observed with microsatellite analysis (**Figure 2a**). STRUCTURE analysis on genetic variants (**Figure 2b**) assigned strains isolated from human faeces to single clusters, indicating a common ancestor with co-clustering strains (Liti et al., 2009; Strope et al., 2015). The only exception was the YA5 strain meiotic segregant (YA5-46A), which could not be assigned to a specific cluster and was hence identified as a mosaic strain, as described in Experimental Procedures (**Figure 2a** and **b**). Ancestry analysis performed on genome-wide data after removal of strain-specific loci (keeping loci with minor allele frequency>0.05 in the set of strains) identified seven populations that were defined as HG1, HG2, HG3, HB (Human Body), WC (Wild Clade), sake, and Lab (Laboratory) (**Figure 2b**). The HG1, HG2, and HG3 clusters were named according to the clustering also observed by means of microsatellite analysis (**Figure 1a**). Three of our identified clusters did not include strains isolated from human faeces: HB (“Human Body”, i.e., strains isolated from various non-gut sites of the human body), WC (“Wild Cluster”, i.e., strains isolated from wild environments and natural fermentations), and the previously identified (Liti et al. 2009) sake cluster (strains used for sake fermentation; see **Figure 2**). The strains isolated from donors E, B, and D were inferred to descend from the HG1 ancestor, previously identified as Wine European (WE) (Liti et al. 2009). As observed with microsatellite data, the strain Y13EU and the meiotic segregants of the strains YUC22, YP4, and YA5 (YUC22-34C, YP4-40D, and YA5-46D respectively) clustered together with the wasp intestine strains YVPC7.6 and BIBVC5.3 descending from the same HG2 ancestor (**Figure 2a** and **b**). The meiotic segregants of the YH1 strain (YH1-28A, YH1-28B, and YH1-28C), clustering with the wasp intestine strain BIBVC1.1 and strains isolated from samples collected in West Africa (from bakery, wine, and other fermentations), were inferred to descend from the HG3 ancestor, previously identified as WA (West African) (Liti et al., 2009). Interestingly, all the strains isolated from clinical samples collected from patients in Italian hospitals (YJM969, YJM975, YJM978, YJM981, YJM984, YJM987, YJM990, YJM993, YJM996 by Strope et al., 2015) grouped in the

HG1 cluster, together with some of the strains isolated from human faeces in this study (YB7, YB8, YB10, YD1, YE2, and YE3, **Supplementary table 1** and in pink in **Figure 2a**). On the contrary, the remaining clinical strains, isolated from hospital located in other countries (Portugal, United Kingdom-Newcastle, Romania, California, District of Columbia, Michigan, North Carolina, Texas-USA), clustered separately from either human faeces or wasp intestine isolates (pink labels in **Figure 2a**).

## Genome-wide investigations on gut strains

To further describe the genome settings of human gut strains, we performed several genome-wide analyses (**Figure 3**) aimed at identifying acquisitions with respect to the reference genome S288c (introgressed CDSs, **Supplementary Table 3**), chromosomal aberrations (through analysis of coverage, **Supplementary Figure 3** and **Supplementary Table 4**), lost/multiplied (see the definition below) coding DNA sequences (CDSs) defined by CDS-based copy number variations (CNVs, **Supplementary Table 5**), and defective CDSs -whose coding sequence showed the insertion of a premature stop codon or to an out-of-frame indel (**Supplementary Table 6**).

In order to discover genes present in newly sequenced strains and missing in the reference strain (referred to as “introgressed”), for each strain, we assembled the reads that did not map to the S288c genome. The total 2,447 identified CDSs were translated and searched in the NCBI non-redundant protein database with BLAST (**Supplementary Table 3**). In order to avoid inflating the relevance of introgressions, we evaluated whether introgressed CDSs could complement (“recover”) lost or defective CDSs (found with the procedure described below) in each strain separately. Recovered CDSs were excluded from the list of introgressed CDSs. Similarly, in the following sections on lost and defective CDSs we refer to lists of CDSs not including recovered genes. A total of 298 CDS-derived proteins proved to confidently derive from previously known proteins (12.1% of 2,455 total newly assembled CDSs). Among these, 233 were derived from other *S. cerevisiae* strains, 30 in other *Saccharomyces* spp. and 15 in *Candida spp.* (**Supplementary Table 3**). Interestingly, some genes derived from other *Saccharomyces spp.* belonged to families such as *FLO* (flocculation), *HXT* (hexose transporters), *PAU* (seripauperin), *ENA* (sodium-pumping

ATPase), *THI* (thiamine synthesis), *IMA* (isomaltase), *PHO* (cell surface glycoprotein involved in phosphate metabolism) (**Supplementary Table 3**). The clustering of strains based on presence/absence of new genes partially reproduced the clustering observed on whole-genome data (**Figure 3a**), the only exception being YUC22-34C, YB7, and YB8 strains that clustered separated from previously co-clustering strains. To investigate whether the presence of new genes could cause gain of functions, we searched for enrichments in specific GO terms (**Supplementary Table 3**). The results showed that the HG2 group was enriched for carbohydrate metabolic process, cation-transporting ATPase activity, hydrolase activity, and flocculation (CDSs from the *FLO* family) (**Supplementary Table 3**). Interestingly, the new CDSs characteristics of the HG1 cluster were enriched in GO terms related to flocculation. In addition, the HG1 group was enriched for NADH dehydrogenase (ubiquinone) activity and respiratory chain (**Supplementary Table 3**).

Coverage analysis carried out on entire chromosomes revealed a few cases of chromosomal polyploidies, namely on chromosome I in the YB7, YB8, and YP4-40D strains, on chromosome IX in the Y13EU, BIBVC5.3, and YVPC7.6 strains, and on chromosome II and the terminal portion of chromosome XI in the YUC22 strain (**Supplementary Figure 3** and **Supplementary Table 4**). It is surprising to observe that the coverage of the above mentioned chromosomes, despite higher, is not above the classic 2x threshold with respect to the rest of the genome. Although this finding could be explained for diploid strains (e.g. YB7, YB8, Y13EU, BIBVC5.3, YVPC7.6, and YUC22), where the given chromosome could be present in three copies and the other chromosomes in two copies, such behaviour in the haploid strain YP4-40D remains elusive. Aiming at the identification of genetic variants associated with and possibly causing the observed polyploidies, we explored the sequences of genes known to be involved in chromosomal segregation, as defined by the Gene Ontology classification. We could not find any genetic variation specific for the strains bearing a polyploid chromosome I (**Supplementary Table 4**). Contrarily, variants in the *BIR1* (Ser825Pro), *CDC45* (Phe362Asn, Gly366Asp, and Gln496Asn), *CHL1* (Asp264Asn and Thr449A), *CHL4* (Asp264Asn and Thr449Ala), *CSM1* (Met178Leu), *DSN1* (Ala133Glu, Asp377Ala, Ser413Cys, and Leu477Stop), *IPL1* (Phe65Cys and Ile140Met), *MCM7* (Thr371Ile), *NDC10* (Leu150Thr, Val179Lys, Pro472Tyr, Asp549Gln), *PLC1* (Trp178Cys and Pro774Ser), *RFC4*

316 (Asp86Gly), *SHP1* (Glu67Gly), *SPC105* (Val827Leu), and *SPT6* (Ser6Leu) genes were specifically
317 associated with a polyploidy in chromosome II (**Supplementary Table 4**). Similarly, the polyploidy
318 in chromosome IX was associated with variants in the *NDC10* (Arg666Asp and Arg666Glu) and
319 *SGO1* (Ala577Gly and Ala577Arg) genes (**Supplementary Table 4**). However, it has to be
320 considered that such results may be affected by the limited amount of analysed strains (e.g. only
321 one strain, YUC22, showed a polyploidy of chromosome II). Strikingly, the polyploidies observed in
322 chromosomes I and IX for strains isolated from human faeces (**Supplementary Table 4**) were
323 previously found to characterize *S. cerevisiae* strains used in Italian industries (Zhu et al. 2016).
324 For each CDS, CNV was calculated as the log2-transformed ratio of the number of reads mapping
325 against the CDS divided by the median coverage of the genome. Notably, the clustering based on
326 the CNV profiles of sequenced strains (**Figure 3b**) generally reproduced the clustering observed
327 through whole-genome sequence comparison (**Figure 2**), with strains being grouped according to
328 their ancestry, even if strains belonging to the HG1 cluster were further subdivided into sub-
329 clusters mainly corresponding to their source of isolation (**Figure 3b**). In order to gain functional
330 insights on CNVs, we defined two categories for each CDS: “lost” CDSs, with median coverage at
331 least twice lower than the median coverage of the whole genome (CNV < -1) and not found among
332 CDSs identified from non-mapping reads (see **Supplementary Figure 4** for a graphical
333 representation) and “multiplied” CDSs, with median coverage at least twice higher than the median
334 coverage of the whole genome (CNV > +1, possibly indicating over-representation compared to the
335 reference strain). A detailed report of this analysis is described in **Supplementary Table 5.** The
336 strains belonging to the HG1 cluster showed the highest number of cluster-specific CNVs, with 47
337 multiplied CDSs (7 of which were Ty elements) and 74 lost CDSs (10 of which were Ty elements).
338 The strains belonging to the HG3 cluster showed only two cluster-specific multiplied CDSs and 44
339 lost CDSs (5 of which were Ty elements). Functional enrichment analysis on the lists of multiplied
340 or lost CDSs specific for either HG clusters or intestinal/non-intestinal strains revealed that HG2
341 and intestinal strains bore a higher copy number of CDSs related to the catabolism of nitrogen
342 sources (e.g. allantoin), as reported in **Table 2** and **Supplementary Table 5.** Interestingly, CDSs
343 multiplied in non-intestinal strains were functionally associated with GO terms related to the plasma

membrane, while CDSs lost in intestinal strains were functionally associated with GO terms related to the cell wall (**Table 2** and **Supplementary Table 5**). Concerning transposable elements, we found a large variability among the CNVs of all the classes of Ty (Ty1, Ty2, Ty3, and Ty4) when comparing strains belonging to the LAB, HG1, HG2, and HG3 clusters, with the sole significant association of lost Ty3 CDSs in intestinal compared to non-intestinal strains (**Supplementary Figure 5**).

We further analysed CDSs having an out-of-frame indel or whose resulting protein after translation contains a premature stop codon (to which we will refer as “defective CDSs”), identifying a total of 414 CDSs defective in at least one of the sequenced strains (**Supplementary Table 6**). A few defective CDSs were found to be replaced by CDSs identified from non-mapping reads and therefore were excluded from further analyses. The clustering of strains’ profiles based on the presence of defective CDSs reproduced the clustering observed through phylogenetic analysis on whole-genome polymorphisms (**Figure 2**), grouping strains according to their ancestry (**Figure 3c**). The location of defective CDSs for each strain in every chromosome is reported in **Supplementary Figure 6.** Similarly to what we did for CDS CNVs, we examined the defective CDSs characteristic of the HG groups by an enrichment analysis on Gene Ontology (GO) terms (using YeastMine tools, fdr<0.05; **Supplementary Table 6**). HG1 and the sequenced strains sharing the HG1 (WE) ancestor were characterized by the highest number of cluster-specific defective CDSs (94 CDSs, **Supplementary Table 6**), enriched in the cell periphery GO term (**Table 2**). Among the 41 CDSs defective specifically in the members of the HG2 cluster, we found an oligopeptide transporter (*OPT1*) and a noncatalytic subunit for phospholipid translocase (*CRF1*) (**Supplementary Table 6**). Finally, exploring the 52 defective CDSs characterizing the HG3 cluster, we found 41 CDSs defective in all these strains of which 7 were related to the cell membrane or wall (*BSC1*, *ECM12*, *FIT2*, *FIT3*, *HVG1*, *MNN5*, and *SUR7*) and 2 were related to cations transportation (*CCC2* and *MMT2*). Interestingly, intestinal strains contained 258 defective CDSs, which were non-defective in strains from other sources and enriched in gene ontologies related to metal homeostasis (**Table 2**). On the other hand, the 37 CDSs defective in non-intestinal strains only did not show any enrichment in gene ontology terms.

Furthermore, we measured the relative frequency of polymorphism in CDSs of sequenced strains and the resulting matrix (**Supplementary figure 7**) was used in a cluster analysis (**Figure 3d**). We found a substantial agreement between the isolates clustering based on frequency of polymorphism (**Figure 3d**) and whole-genome phylogenetic comparison (**Figure 2**).

## Phenotyping of gut strains

Considering the relevant genetic features varying between gut strains and among strains isolated from faecal samples and from other sources, we asked whether these differences were reflected at the phenotypic level. Hence, we studied gut strains for differences in traits relevant for growth and survival in the gut environment: invasiveness, resistance to several physiological temperatures and pH, and sporulation (McCusker et al., 1994; Diezmann and Dietrich 2009) (**Figure 4** and **Supplementary Table 7**). Additionally, we evaluated whether the identified populations subtended phenotypic differences in the *S. cerevisiae* isolates (**Figure 4** and **Supplementary Table 7**). The YH strains, belonging to HG3 cluster, were the only human intestine strains able to form pseudohyphae and resist to oxidative stress and showed the highest sporulation rate (>30% asci/cell at 37°C; **Figure 4** and **Supplementary Table 7**). HG1 strains (especially YD, YE, and some YB) were unable to invade the medium and sporulate (**Figure 4** and **Supplementary Table 7**).

While all the tested strains showed the same ability to grow at various pH values, a wider variability was observed in the growth ability at various temperatures among the tested strains (**Figure 4**). We found that the growth abilities at 4°C, 40°C, 42°C and 44°C were dependent on the intestinal origin of the strains, with strains isolated from intestines (either human or insect) having a slightly increased fitness at 4°C (Chi-squared test $p=0.022$) and a decreased fitness at 40°C (Chi-squared test $p<0.001$), 42°C (Chi-squared test $p<0.001$) and 44°C (Chi-squared test $p<0.001$) compared to strains isolated from other sources (**Figure 4**). Furthermore, intestinal strains (isolated from either human faeces or insect intestines) showed a significantly lower sporulation efficiency at 28°C and 37°C than strains isolated from other sources (Mann-Whitney U test $p<0.001$, **Supplementary figure 8**). On the other hand, a few traits statistically differed between human and wasp intestinal

strains: human gut isolates showed a higher fitness at 40°C (Chi-squared test p<0.001) and at 42°C (Chi-squared test p=0.001) than wasp isolates (**Figure 4**).

Interestingly, the gut strains' sporulation efficiency mirrored the microsatellite-based clustering and whole-genome based ancestral lineages (**Figure 1a** and **Figure 2**). Indeed, the HG1 group, deriving from the HG1 (WE) ancestor showed low sporulation rate (0% or ranging from 10% to 15%), the HG2 group showed medium/low sporulation rate, and HG3 strains (with HG3 (WA) ancestry) showed the highest percentage of sporulation (>30%; **Supplementary Table 7**). The sequence variation of the *RME1* gene, a negative regulator of sporulation, has been shown to correlate with the strains' sporulation efficiency (Gerke et al., 2009). In particular, the insertion of an adenine in position -308 (upstream to the gene's codon start) was found to be associated with an increase in the strain sporulation efficiency (Deutschbauer and Davis 2005). We confirmed the relations between the adenine insertion and the sporulation efficiency for our strains (**Supplementary Figure 9**). In addition, according to our results, the clustering of *RME1* sequences (also including an upstream region encompassing the previously identified relevant locus, **Supplementary Figure 9**) reproduced the clustering obtained with microsatellite analysis (**Figure 1a**), and depicted the same strain grouping identified by means of ancestry analysis (**Figure 2**).

Since genome sequencing indicated a genetic variation in cell wall and mannose genes as well as sugar metabolism genes in HG strains, we analysed the cell wall sugar composition (see Experimental Procedures) of a selection of 18 strains isolated from human faeces or other sources. HPLC analyses revealed that HG strains' cell wall shows a significantly lower content of mannose compared to laboratory and grape strains, a lower content of glucose compared to grape strains (Mann-Whitney U test, fdr<0.05, **Figure 5a**) and a higher content of galactose compared to grape strains (fdr<0.01; **Figure 5a**). In addition, strains isolated from human faeces show a higher content of glucosamine in their cell walls compared to strains isolated from all the other tested sources (grapes, laboratory, and insect intestines; fdr<0.05; **Figure 5a**). Furthermore, HG1 strains showed a higher mannose content in their wall compared to strains deriving from the laboratory ancestor (fdr<0.05), and significantly lower amounts of galactose compared to strains deriving from

both HG2 and laboratory ancestors (**Figure 5b**, fdr<0.05).

## Immunophenotyping of *S. cerevisiae* isolates

As observed through analyses carried out on whole-genome sequences and confirmed by cell wall composition analysis, human gut strains show a characteristic cell wall composition. Considering that the sugar moieties of cell wall are the principal antigens sensed by the host immune system during host-fungal interaction (Perez-García et al., 2011; Lewis et al., 2012; O'Meara et al., 2015), we evaluated whether the observed differences in cell wall composition and the genomic traits were mirrored by the strains' immunogenicity. Direct contact of immune cells with fungi leads to activation of antigen presenting cells and subsequent priming of adequate pro-inflammatory T helper (Th) response or regulatory T cells (Treg) response (Romani 2011). Previous studies showed a fungal strain-dependent variation of immune recognition and consequent immune reactivity depending on strain features (Marakalala et al., 2013; Rizzetto et al., 2013; Smith et al., 2014; Cavalieri et al., 2018). Hence, to characterize the immune response elicited by human gut *S. cerevisiae* isolates, peripheral blood mononuclear cells (PBMCs) were challenged with HG isolates. Strains isolated from grapes (Sgu52 and Sgu421), natural environments (BB1533 and BT2240) and a laboratory strain (SK1), were used as controls. The induced cytokine profiles were determined by using the Human Milliplex® assay (Merck-Millipore) (**Supplementary Figure 10** and **Figure 6a-c**). We measured the pro-inflammatory cytokines IL-6, IL-1β and TNF-α, expressed by antigen presenting cells in response to fungal antigens (Qin et al., 2016; van de Veerdonk et al., 2017), as well as the Th response cytokines IFN-γ (Th1) and IL-17 (Th17), both known to allow resistance towards colonization of fungi such as *Aspergillus fumigatus* (van de Veerdonk et al., 2017) and *Candida albicans* (Gaffen et al., 2011; Gozalbo et al., 2014) and the Th2 cytokine IL-13, involved in allergic response (Gagliani and Huber 2017). We also measured the level of IL-10, an anti-inflammatory cytokine involved in immune tolerance towards fungi (Roussey et al., 2016). The *in vitro* assay showed strain-specific differences in the cytokine profiles triggered by the challenge (**Supplementary Figure 10** and **Figure 6a-c**) and revealed the Th-polarizing cytokines IL-17A, IFN-γ, and IL-10 as the main factors discriminating HG strains induced- immune responses

(**Figure 6 and Supplementary Figure 10**). The variable cytokine profiles induced by the environmental strains and human faeces isolates indicates strain-specific rather than source specific immune responses (**Figure 6d** and **Supplementary figure 4**). The Principal Component Analysis carried out on quantified IL-17A, IFN-γ, and IL-10, and on strain sporulation levels reproduced the same strain grouping (**Figure 6d**) previously observed by comparing the strains' whole-genome sequences (**Figure 2**). Non-sporulating HG1 strains (i.e. YE1, YB10, YD1) showed a tendency towards the induction of IFN-γ-mediated responses. HG2 strains with low sporulation efficiency (i.e. YA5, YB7, YP4 and YUC22) induced high IL-17A production (**Figure 6d**). The opposite trend was observed for the high sporulator YH1 and SK1 strains, derived from the HG3 (WA) cluster, in which the high IFN-γ mediated-inflammatory response was counterbalanced by high levels of the immunosuppressive cytokine IL-10 (**Figure 6d**).

# Discussion

In this work we performed genome-scale and phenotypic analyses aimed at exploring the possible features of *S. cerevisiae* enabling its survival and possible inhabit in the human gut, therefore candidating it to enter the list of the much better documented natural (e.g. grape, soil, trees) or artificial (e.g. fermentation) yeast environments.

It is known that a few fungal species, mostly *Candida spp.*, are able of growing and colonizing the gastrointestinal tract. On the other hand, strains from species considered like "passengers", rather than colonizers, such as *S. cerevisiae,* have been shown to impact gut ecology (Hallen-Adams and Suhr 2017), and Ott and collaborators reported an enrichment in IBD patients' faecal samples of *S. cerevisiae* as well as other fungal species (Ott et al., 2008). Although obtained in a limited number of donors and isolated strains, we found that *S. cerevisiae* is more abundant in CD patients than healthy controls. Microsatellite analysis revealed that gut strains show genetic similarities with foodborne strains (HG1 and HG2 clusters), but also with strains isolated from other sources collected in distant geographical locations, such as West Africa (HG3 cluster). The hypothesis that some of the human intestine strains are foodborne is also supported by the analysis of

chromosome poly-ploidies. In fact, poly-ploidies in chromosomes I and IX were found to characterize Italian industrial strains (Zhu et al. 2016), and were also found in the present study in some of the human intestine isolates descending from the HG1 and HG2 ancestors.

Notably, strains isolated from the same donor were more similar among each other than strains sharing the same genetic cluster or randomly sampled from the strain dataset. These indications suggest that either the strains isolated from faeces of the same patient arose from clonal expansion within the gut or that patients were colonized by clonal strains from the same source. Either way, the observed residence of *S. cerevisiae* clonal populations only in patient intestines may indicate that a gut microbial alteration (dysbiosis) occurring in chronic inflammation could admit the expansion or the survival of strains showing particular characteristics (Koh 2013). Indeed, increase of the intestinal permeability, dysbiosis or even impairment of the immune system are all conditions previously documented as complicating the fungal-host interplay (Underhill and Iliev 2014; Chiaro et al., 2017). The comparison of human gut isolates with strains isolated from wasp intestines revealed further insights on this topic. Wasp intestines isolates were previously shown to represent the regional genetic and phenotypic variability of *S. cerevisiae* (Stefanini et al., 2012; Dapporto et al., 2016). In the current study we report that some wasp intestine isolates cluster together with human gut isolates, indicating either that humans are exposed to (and colonized by) a limited portion of the regional *S. cerevisiae* strains or that only a few strains are able to persist/survive in the human intestine. It is intriguing to observe that among all clinical strains isolated in other studies, only those isolated from patients in Italian hospitals (Liti et al. 2009; Strope et al., 2015) clustered together with a few strains isolated from human faeces (in the cluster HG1), suggesting that environmental and foodborne strains from the same geographical region can be the source to select clinically-associated yeasts, in individuals predisposed to be colonized by them. Conversely, human gut isolates also encompassed strains belonging to other clusters, partially reproducing what previously observed for wasp strains (Stefanini et al., 2012). Nevertheless, wasp strains encompass a wider genetic variability than human gut strains, again suggesting that the latter may be subjected to selection (Stefanini et al., 2012). Prompted by the observations on microsatellite data, we searched for genetic characteristics eventually

characterizing gut strains by means of whole-genome sequencing. CNV analysis revealed shared peculiarities for strains isolated from human and insect guts, compared to strains isolated from other sources (grapes and wine). In particular, intestinal strains showed a higher number of copies of CDSs involved in allantoin metabolism, and a lower number of copies of CDSs related to the vacuole function and to the cell wall compared to non-gut strains. The modulation of such function, possibly caused by the modification of CDS numbers, may result in a physiological advantage for strains in the intestinal environment. In fact, high allantoin levels have been reported in the body fluids of IBD patients, resulting from the general inflammation status associated with the disease (Schicho et al., 2012). Differently, the complete profile of either lost or new CDSs did not differentiate the strains according to the isolation source, but helped in further understanding the genetic differences among strains isolated from faecal samples.

Considering the relevant genetic features varying among intestinal strains and between strains isolated from faecal samples and from other sources, we wondered whether these differences were reflected at the phenotypic level. The characterization of strains isolated from intestines for phenotypes putatively relevant for microbial survival in the human gut revealed that strain sporulation efficiency and sequence variation in a specific sporulation-associated allele of *RME1* (Gerke et al., 2009) mirrored the strains grouping observed by comparing the whole-genome sequences of intestinal strains, mostly for the HG3 cluster composed by highly sporulating strains. This result suggests a connection between sporulation and ability of certain strains to survive in the human (inflamed/dysbiotic) gut. Furthermore, wasp and human intestine isolates share an increased fitness at 4°C, while most of them have a reduced growth rate at 40°C, 42°C and 44°C, and a generally lower sporulation efficiency at 28°C and 37°C (absent in HG1 cluster) compared to strains isolated from other sources. These traits specific for several intestinal strains may suggest that *S. cerevisiae* is allowed to reside and grow within the intestine, provided that it maintains slow growth and low sporulation rates. We may hypothesize these strains characterized by a moderate division and sporulation rates may have higher chances to persist/survive within the human intestine, in particular in presence of the complex inflammatory environment such as that observed in CD patients (Brand 2009). Healthy human immune cells have been shown to mount a Th1

inflammatory response - counterbalanced by a Treg response - against *S. cerevisiae* cells through the recognition of cell wall mannans (Rizzetto et al. 2010). Conversely, lack of exposed mannans led to the induction of a Th17 inflammatory response to spores, enriched in chitin, a molecule exposed in *S. cerevisiae* bud scars and in spore cell wall (Rizzetto et al., 2010). It is interesting that previous metagenomics studies in CD patients (Liguori et al., 2016) showed that an enrichment in *S. cerevisiae* is associated with decrease of *Candida* and with remission. Thus, we could speculate that *S. cerevisiae* populations colonizing the gut use the host's immune system to fight potential competitors, and that this reduction in opportunistic pathogens (such as *Candida*) is benign in pathological conditions, such as CD.

Probably even more strikingly, human intestine isolates showed a peculiar composition of their cell wall, bearing in particular a higher content of galactose and glucosamine compared to the strains isolated from all the other sources. The cell wall enrichment in galactose residues has been associated to filamentous growth, cell-cell adhesion, flocculation in *S. pombe* (Tanaka et al., 1999) and biofilm production in *C. albicans* (Cavalieri et al. 2018), favouring invasiveness, further indicating that gut colonization requires specific traits. Thus, the peculiar cell wall composition of these strains appears to be a gut-specific feature and could represent a selective advantage to survive and expand in the host gut. Further and more specialized studies are necessary to evaluate the role of galactose in the strains persistence in the gut.

Concerning the strains' immunogenicity, only strains having high sporulation efficiency or belonging to the HG3 cluster induce a concordant immune response, triggering a mixed response (the high level of pro-inflammatory IFN-γ-mediated-Th1 response was counterbalanced by strong IL-10 induction) potentially enabling yeasts to escape immune-surveillance. On the contrary, the other two classes of human gut strains (non- and low-sporulating or belonging to HG1 and HG2 clusters) triggered elusive and apparently incoherent immune responses. Overall, the observations highlight a remarkable correlation between yeast isolates' specific genetic make-up and immunogenicity, suggesting alternative but convergent strategies for the colonization of the gut environment.

Despite the results reported here should be validated in larger cohort of faecal donors and yeast isolates, the findings indicate how, from an evolutionary perspective, the gut environment could

serve as a reservoir and evolutionary niche for *S. cerevisiae*, in which genetic makeup could ultimately influence its immunogenicity, highlighting the need to consider the interplay between fungal cell wall and gut immune function in determining yeast population selection. Our observations highlight three conditions correlating with a *S. cerevisiae* strain ability to survive and persist in the human intestine: *i* - the ability to exploit the harsh environment, by metabolizing allantoin, present at high levels in IBD patients due to the action of inflammation-associated high radical oxygen species on uric acid (Schicho et al., 2012); *ii* - a cell wall enriched in mannose, possibly allowing their persistence in the pre-existent Th1-based inflammatory environment of CD patients' intestines through IL-10 production (Brand 2009) and iii) a moderate sporulation efficiency and growth rate at temperatures characterizing inflamed intestines, avoiding the induction of a Th17-driven inflammatory immune response able to kill/eradicate yeast cells (Rizzetto et al. 2010, Reuter, et al. Curr Biol. 2007)*.*

In intestines characterized by an inflammatory condition, as we observed in CD patients, the residence of strains with peculiar genetic characteristics and cell wall composition could be the result of selection and adaptation to a peculiar gut environment. In this perspective, *S. cerevisiae* is Generally Recognized As Safe (GRAS), since its ability to induce trained immunity provides the host with benefits such as improved immunity and resistance to pathogens. This could lead us to reconsider the definition of "domestication" of *S. cerevisiae*, emphasizing that humans, wasps, and possibly all food-associated insects or other animals, could have vectored *S. cerevisiae* among fermented food and beverages. Studying the evolution of the strategies used by yeast to evade immune surveillance or to provide a selective advantage to the host by modulating immune responses could lead to discovery the genetic features potentially turning a friend into a foe. In this novel perspective, the combination of genotyping, phenotyping, and immunophenotyping approaches described in this paper is a powerful approach to describe *S. cerevisiae* diversity.

# Experimental Procedures

## Strain isolation, identification and typing

A total of 93 paediatric subjects were enrolled at the Meyer Children's Hospital (Florence, Italy). All parents or caregivers of paediatric faecal donors (healthy or IBD patients) were made aware of the nature of the experiment, and gave written informed consent for faecal sample collection, in accordance with the sampling protocol approved by the Ethical Committees of the Meyer Children's Hospital and the Azienda Ospedaliera Universitaria Careggi, Florence, Italy (Ref. n. 87/10). All individuals were Caucasian and their age ranged from 4-19 years. Faeces were collected from all paediatric subjects by physicians and caregivers who had been instructed to collect faecal samples under sterile conditions to avoid contamination. A 1 ml faeces aliquot was plated onto YPD agar medium (1% yeast extract, 2% peptone, 2% D-glucose, 2% agar) supplemented with chloramphenicol (1 mg/ml) to avoid bacterial growth, and incubated for 2-3 days at 28° C. Yeast genomic DNA was extracted as previously described (Hoffman and Winston 1987). Strains were identified by sequencing the ITS1-5.8S-ITS2 region, using the ITS1 and ITS4 primers as previously described (**Supplementary Table 2**) (Sebastiani et al., 2002). *Saccharomyces cerevisiae* strains were identified by ITS1-5.8S-ITS2 region sequence homology with the *S. cerevisiae* ITS1-5.8S-ITS2 region sequences available in the NCBI database. The strains identified as *Saccharomyces cerevisiae* were used in following analyses and compared to strains isolated from different sources. For strain typing, a set of 286 strains isolated from various sources (**Supplementary Table 1**) was characterized in terms of allelic variation at 12 microsatellite loci (Legras et al., 2007). The 12 microsatellite loci were amplified as previously described and using the primers listed in **Supplementary Table 7**. The PCR amplicon sizes of the 12 loci were assessed by capillary electrophoresis. The chord distance Dc matrix was calculated for each strain with a laboratory-made program. The phylogenetic tree was obtained with the Neighbor-joining method from the distance matrices with the Phylip Neighbor 3.67 package and drawn up using MEGA4.0 (Tamura 2011). The tree was rooted using the midpoint method. Strain clustering was investigated by means of K-means analysis on chord distances, with the best number of clusters being identified as the lowest number of clusters minimizing the distances

among strains of the same cluster. Significant differences among distances between strains grouped according to faecal donor or clustering were evaluated by means of Mann-Whitney U test as previously done (Novitsky et al., 1999; Gilbert et al., 2001). In addition, distances among strains isolated from the same donor were compared with distances among an equal number of strains randomly sampled from the set of strains characterized by means of microsatellite analysis. Strains from healthy subjects (Y13EU and YN19) and from patient A (YA5) were not included in the comparison because a single strain was isolated from the corresponding faecal donor. 10,000 sampling iterations were carried out for each comparison, for a total of 60,000 comparisons (10,000 iterations x 6 groups of strains isolated from the same faecal donor). P-values were corrected for multiple testing (false discovery rate, fdr), and differences were considered significant if fdr<0.05.

## Genome sequencing, reads alignment and genotype calling

For whole genome sequencing purposes, strains were selected to maximize the genomic variety inferred from microsatellite phylogeny. Strains whose genome were sequenced in this study are listed in **Supplementary Table 1.** Whole genome sequencing was performed using the Genome Analyzer IIx (GAIIx) platform and HiSeq2000 sequencing instruments. The standard Illumina protocol with minor modifications was followed for the creation of short-insert paired-end libraries (Illumina Inc., Cat. # PE-930-1001), as reported in Cavalieri et al. 2018.

Illumina run generated 16,504,557 ± 7,966,339.6 quality paired end sequences, with a read length average of 113.8 ± 21.4 base pairs (excluding the primer sequences). On average, we found ~16.5 million filtered reads and 14.9 million mapped reads. Illumina reads were subjected to quality control (filtering and trimming) using NGS QC Toolkit v2.3.3 (Patel and Jain 2012). Paired reads were filtered with parameters -l 70, (cutOffReadLen4HQ) and -s 20 (cutOffQualScore), allowing the elimination of reads with a PHRED quality score lower than 20 for more than 30% of their lengths. Moreover, reads were trimmed at the 3' end for bases with a PHRED quality score lower than 30. Paired reads were mapped to the reference genome (*Saccharomyces cerevisiae* S288c, NCBI ID=559292) using Burrows-Wheeler Aligner (BWA 0.7.12) (Li and Durbin 2009).

The Genome Analysis Toolkit (GATK 2.1) was used for base quality score recalibration, Indel (insertion or deletion) realignment, duplicate removal, and to perform SNP and Indel discovery (McKenna et al., 2010), resulting in strain-specific vcf files. The coding sequences of the investigated *S. cerevisiae* strains were generated using variant imposition (a.k.a. consensus calling) in order to create uniform sets of genes for further analysis. Briefly, this technique inserted variants (SNPs and Indels) produced by GATK and specific for each strain into the coding sequences of the reference strain (**Supplementary Table 2**) to create accurate CDSs useful for phylogenetic comparison across strains. Other isolates previously published and used in the comparison (listed in **Supplementary Table 1**) were 1) identified in the NCBI Assembly database and their CDSs ("CDS from genomic" file type) downloaded from GenBank using NCBI Download Assembly tool (strains genomes published by Strope and collaborators (Strope et al., 2015)) or 2) Sanger Institute FTP repository (cere_assemblies.tgz) for strains published by Liti and collaborators (Liti et al., 2009). Other strains available in NCBI Assembly were evaluated for their degree of CDS annotation and judged not suitable for inclusion. However, we considered this set of strains sufficiently broad, in terms of number of strains, and considering their environmental isolation sources and genetic variability, to allow an exhaustive comparison of *S. cerevisiae* populations. As a whole, the dataset of genomic sequences encompassed 29 newly sequenced strains and 119 strains previously described: 38 from clinical samples (collected from various parts of the human body, usually in presence of co-occurring infections) and 81 from various sources; such strains were sequenced in previous studies: 37 from the Liti et al. study (Liti et al., 2009), 82 from the Strope et al. study (Strope et al., 2015).

For ortholog identification, fasta headers of sequences were parsed using the "gene=" tag as an annotation key. CDSs described as unknown or not-annotated, absent from the reference strain S288C, or pertaining to mitochondrial or plasmid DNA were excluded from further analyses. All the sequencing data generated for this study are publically available both as raw and processed whole-genome sequences (ENA ID: ERP002541 and Array Express ID: E-SYBR-8).

## Phylogenomic analysis of CDS

The coding sequences of genes present in the genomes of the previously described 144 selected *S. cerevisiae* strains (see the “Genome sequencing, reads alignment and genotype calling” paragraph) were collected in gene-specific sequence files in order to produce a coherent collection of orthologs. The resulting 2,297 ortholog sets were individually processed by reverse multiple sequence alignment (rMSA). Briefly, CDSs were translated, aligned with clustal-omega (Sievers and Higgins 2014) and then codons were juxtaposed in coding sequences using protein alignment as a guide. This allowed the creation of CDS-based ortholog specific MSAs, that were eventually reduced in size (giving a “reduced MSA”) by removing conserved alignment columns. In order to avoid inflating differences among strains due to aberrant alignments (usually due to frameshifting indels in at least one strain, leading to a gap rich alignment) all orthologs whose reduced MSA was longer than 50% or the initial MSA, were considered unreliably aligned and excluded from further analyses. This procedure allowed us to evaluate whether CDSs with potentially confounding alignments were truly polymorphic or aberrant. In fact, while polymorphic genes are relevant to disclose species evolution, as they produce correct alignments and maintain *bona fide* a role in selective pressure, aberrant mutations may not be informative because producing inconsistent alignments and reasonably ignored by selective pressure due to a loss-of-function of the defective gene.

The remaining reduced ortholog MSAs were concatenated to build a single CDS-derived “genome-wide” alignment, as previously described (Ramazzotti et al., 2012) involving 144 strains and encompassing 1,715 genes, for a total of 245,990 informative loci. This alignment was eventually used to produce a phylogenomic tree using FastTree 2.1.10 SSE3 (Price et al., 2010) with default parameters and for population structure analysis.

## Population structure analysis

For investigations of population structure, the previously described CDS-derived “genome-wide” alignment was analysed by using fast*Structure* to assign individual strains to populations (Raj et al., 2014). Aiming at the identification of the best number of populations (K) in our dataset, we first binned SNPs by minor allele frequency (maf, only loci with maf > 0.05 were used for this analysis),

then fastStructure analysis was carried out testing from 3 to 13 K values (Raj et al., 2014). The best K was identified by using the chooseK.py script of the fastStructure software (Raj et al., 2014). Finally, strains were assigned to a given population if showing at least 60% ancestry from that population. Strains assigned to the same cluster are supposed to descend from the same ancestor. Strains with less than 60% ancestry from any of the other populations were considered mosaics. Despite strains not assigned to a given cluster are commonly considered mosaics (e.g. (Liti et al. 2009) and (Knight et al., 2015), we acknowledge that they could potentially belong to cluster currently not represented among the *S. cerevisiae* strains included in our dataset.

## Identification and analysis of S288C introgressions

In order to discover the CDSs present in newly sequenced strains and missing in the reference strain S288C (previously termed as introgressions), the pool of Illumina reads failing to align onto the reference strain genome was assembled using ABySS 1.9.0 (Simpson et al., 2009), adjusting the k-mer value so that contigs with the highest N50 were maximized. The AUGUSTUS 2.5.5 gene finder (Stanke et al., 2008) was then run on the resulting assembled contigs using the default *Saccharomyces cerevisiae* gene model. The identified CDSs were translated and searched with Protein-Protein BLAST 2.2.31+ against the Non-redundant RefSeq protein dataset. NCBI BLAST was used to annotate CDSs, considering the best hit as that presenting the highest value of the ranking score S = alignment length/protein length * identity %, therefore favouring hits that spanned the whole query protein. In order to avoid inflating the relevance of introgressed CDSs, we further evaluated whether introgressed genes could complement (“recover”) lost or defective CDSs in each strain separately. Briefly, the original S288c CDSs deemed as lost or defective in a specific strain were searched (BLAST) against the pool of introgressed CDSs from this strain, filtering hits using the score described above. Introgressions with a score >90 for a lost or defective CDS were considered “recovery” events, with the effect of removing recovered CDSs from the matrices of defective and lost CDSs described in next sections. Truly introgressed CDSs (i.e. those not recovering original genes) were eventually collected and further analysed as a presence/absence binary matrix (**Supplementary table 3**) involving all strains. Complete

hierarchical clustering on the binary matrix was performed using the hclust function of the R statistical package on a Jaccard's Index-based distance matrix from the R package Vegan (Dixon 2003). For the functional enrichment analysis of introgressed gene lists selected on the basis of strain clustering NCBI IDs were converted in UniProt codes, then a Fisher Exact Test was performed against a reference pan-genome (obtained by combining S288C proteins and all the introgressed proteins, and GO-annotated by EBI quickGO), with p-values adjusted for multiple comparisons using Benjamini-Hochberg correction. Significantly enriched Gene Ontology terms were identified as having p-value<0.05.

## Coverage analysis and CDS-copy number estimation

Starting from bam files created by BWA and optimized with GATK (see above), the samtools software (mpileup command) was used to extract the coverage value for all reference positions in the different strains. For each strain analysed, the median coverage of the entire genome *M* as well as the median coverage of specific CDSs *m* (coordinates were extracted from the corresponding gff annotation file) was computed. The value $CNV = \log2(m/M)$ was calculated for each CDS, representing the copy number variation (CNV), i.e. the excess or defect in coverage with respect to the sequenced genome. This allowed direct comparison of CNV within the entire set of newly sequenced genomes. In the text, we refer to genes having CNV lower than -1 as “lost” (i.e. their CDS median coverage is at least twice lower than that of the rest of the genome). Recovered CDSs (see details in paragraph “Identification and analysis of S288C introgressions”) were removed from the lists of lost CDSs. CDSs were identified as cluster- or source-specific if their CNV was lower than -1 or higher than 1 in at least a strain of the given group and not in any strain of the other groups (**Supplementary Table S3**). To graphically summarise the distribution of CNVs along the chromosomes (and generate **Figure 3a**), the average CNV values, not including TY elements nor subtelomeric CDSs, along each chromosome in non-overlapping bins sized 5000bp were calculated for strains grouped according to their clustering. Functional enrichment (i.e., Gene Ontology enrichment on molecular function) was carried out with YeastMine (Balakrishnan et al., 2012) on lists of CDSs grouped according to either strain clusters or isolation sources; enrichment

significance was tested by means of hypergeometric test and corrected through Holm-Bonferroni approach, with significantly enriched Gene Ontology terms identified as having p-value<0.05. Complete hierarchical clustering on the binary matrix encompassing the CNVs for all genes and all strains was performed using the hclust function of the R statistical package on a Jaccard's Index-based distance matrix from the R package Vegan (Dixon 2003).

To evaluate the presence of chromosomal aneuploidies we used, for each chromosome, the CNV formula specified above on non-overlapping genome stretches of 5000 bp, irrespective on the presence of CDSs. Statistically supported polyploid chromosomes were identified by applying the same approach adopted by Zhu and collaborators (Zhu et al., G3, 2016). Briefly, for each strain we calculated the difference of the median coverages between every couple of chromosomes. The significance of this difference was evaluated by applying a Mann-Whitney U-test between the median coverages, calculated in 5000bp bins, of the two compared chromosomes. The p-values were then corrected for multiple testing (fdr). If the absolute value of the difference of coverage between two chromosomes was higher than the 35% of one of the two compared chromosomes, and the comparison had an fdr<0.05, the two chromosomes were considered as having different ploidies. Once identified the strains showing consistent aneuploidies (e.g. a given chromosome had a coverage significantly differing from the large majority of the chromosomes of the same strain), we searched for associations between aneuploidies and genetic variations in genes known to be associated with chromosomal segregation. Aiming at this, we obtained the list of genes annotated as belonging to the Gene Ontology “chromosomal segregation” (SGD) and identified the genetic variants for each strain. Then, we grouped strains as presenting polyploidies or having specific polyploid chromosome and we searched for genetic variants specific for each strain group.

## Identification and analysis of defective S288C genes

The CDSs obtained by consensus calling in each strain were systematically evaluated in order to obtain information about the presence of possible detrimental mutations that can affect the functionality of encoded proteins. In particular, we considered as highly detrimental both the presence of an out-of-frame indel as well as the presence of premature stop codons. CDS

presenting such conditions were defined as “defective”. Recovered CDSs (see details in paragraph “Identification and analysis of S288C introgressions”) were removed from the lists of defective CDSs. The list of remaining defective genes was further analysed as a present/absent binary matrix (**Supplementary Table 5**) involving all strains and S288C genes. Complete hierarchical clustering on the binary matrix was performed using the hclust function of the R statistical package on a Jaccard's Index-based distance matrix from the R package Vegan (Dixon 2003). To graphically represent the distribution of defective genes along the chromosomes, such genes were counted along each chromosome (in window sized 5000bp) for strains grouped according to their clustering. Functional enrichment was carried out with YeastMine (Balakrishnan et al., 2012) on lists of CDSs, not including TY elements and subtelomeric CDSs, characteristics for either clusters or isolation sources. Enrichment significance was tested by means of the hypergeometric test and the resulting p-value corrected through the Holm-Bonferroni approach, with significantly enriched Gene Ontology terms identified as having an adjusted p-value<0.05.

## Analysis of polymorphism frequency in CDS

The polymorphism frequency of each CDS was calculated for each whole-genome sequenced strain from strain specific vcf files (see previous paragraphs) as the ratio of the number of SNPs in a gene and its bp length (as in the S288c reference, **Supplementary Table 2**). The relative polymorphism frequency (RPF) estimator was designed as a functional score correlated with the probability of maintaining the original function of the CDS: the higher the score, the lower the possibility that the gene function is maintained, finally mimicking CDS deletion and affecting the relative pathway. The complete matrix of polymorphism frequencies was investigated by complete hierarchical clustering using chord distance as a distance metric.

## Phenotyping and immunophenotyping of *S. cerevisiae* strains

Each strain isolated in this study was characterized for the following phenotypes: invasiveness, pseudohyphal formation, growth at supra optimal temperatures, sporulation rate, and pH impact on growth, all performed as previously reported (Strati et al., 2016b). Growth was qualitatively scored,

sporulation efficiency was calculated as the number of tetrads on the total of counted cells. Differences among the qualitative phenotypes of strains grouped according to either the genetic cluster or the isolation source (intestinal or non-intestinal) were evaluated through Chi-squared test of independence by using the chisq.test() function of the MASS R package (Venables and Ripley 2002). Differences among quantitative phenotypes were evaluated through Mann-Whitney U test. Additionally, cell wall sugar composition was assessed through cell wall extraction and sugar quantification (glucose for glucan content determination, mannose and glucosamine for chitin content) by using high-performance anion-exchange chromatography coupled with pulsed electrochemical detection (HPAEC-PAD), as previously described (Cavalieri et al., 2018). Aiming at the characterization of the immune response elicited by the human gut isolates, peripheral blood mononuclear cells (PBMCs) were challenged with *S. cerevisiae* cells. PBMCs were isolated from buffy coat obtained from 6 healthy donors (age > 18 years old) according to the protocol approved by the Ethical Committee of the Azienda Ospedaliera Universitaria Careggi, Florence, Italy (Ref. n. 87/10). Written informed consent was obtained for all the blood donors. Blood separation, cell culture and stimulation were performed as previously described (Cavalieri et al., 2018). Challenges were carried out by culturing PBMCs ($10^5$ cells/ml) with or without (negative control) live yeasts ($10^6$ cells/ml) in RPMI 1640 medium supplemented with 10% heat-inactivated foetal bovine serum, 2 mM L-glutamine, at 37 °C under 5% $CO_2$. At the time indicated below, supernatants were collected and stored at -20°C until assayed. Cytokine detection was performed using the Human Milliplex® assay (Merck-Millipore) according to the manufacturer's instructions using Luminex technology. TNF-α, IL-1β, and IL-6, were detected in 24h supernatants; IFN-γ, IL-10, IL-13 and IL-17A in 5 days supernatants.

## ***RME1*** **gene sequencing and analysis**

For each strain isolated in this study, the *RME1* gene and 599 nucleotides upstream the corresponding start codon were sequenced through Sanger sequencing using the RME1_FWseq and RME1_RVseq primers listed in **Supplementary Table S7**. As a reference, other sequences of this gene were obtained from the NCBI database or by extracting the region of interest from

genomic sequences downloaded from public databases as described in previous paragraphs. Sequences were aligned using clustalW and the SNP sequences were used to calculate the Kimura-2 parameters distances used to generate the neighbor-joining tree.


## Acknowledgements

The authors would like to thank the CNAG-CRG, Centre for Genomic Regulation (CRG) (Barcelona Institute of Science and Technology, BIST, Barcelona, Spain), and in particular Ivo G. Gut, Marta Gut, and Simon Heath, for carrying out whole-genome sequencing, Jean-Luc Legras for microsatellite analysis, Manuel A. S. Santos for supplying natural Portuguese yeast strains, Andrea Morandi for critically reading the manuscript, and Misha Kapushesky for support in whole-genome sequencing analysis.

This project was supported by European Union's Seventh Framework Programme [FP7/2007-2013] under grant agreement n° HEALTH-2010-242220 ("SYBARIS"), by Progetto "NUTRA-TOSCAFRICA" (No.50), Regione Toscana Bando Nutraceutica (DD650-2014), by Progetto giovani si, POR FSE 2014-2020, "VESPATER" project, Regione Toscana, and by Ente Cassa di Risparmio di Firenze Grant 0875. The authors would also thank the Foundation "Amici onlus associazione, malattie infiammatorie croniche intestinali" for support.


## Conflicts of interest

The authors declare that they have no conflicts of interest with the content of this article.

# References


- Aa, E., Townsend, J.P., Adams, R.I., Nielsen, K.M., Taylor, J.W. (2006) Population structure and gene evolution in *Saccharomyces cerevisiae*. FEMS Yeast Res. 6:702–15.
- Almeida, P., Barbosa, R., Bensasson, D., Gonçalves, P., Sampaio, J.P. (2017) Adaptive divergence in wine yeasts and their wild relatives suggests a prominent role for introgressions and rapid evolution at noncoding sites. Mol. Ecol. 26:2167–82.
- Angebault, C., Djossou, F., Abélanet, S., Permal, E., Ben Soltana, M., Diancourt, L., *et al*. (2013) *Candida albicans* is not always the preferential yeast colonizing humans: a study in Wayampi Amerindians. J. Infect. Dis. 208:1705–16.

- Balakrishnan, R., Park, J., Karra, K., Hitz, B.C., Binkley, G., Hong, E.L., *et al.* (2012) YeastMine-an integrated data warehouse for *Saccharomyces cerevisiae* data as a multipurpose tool-kit. Database 2012:bar062-bar062.
- Benjamini, Y., Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. 57:289–300.
- Boix-Amorós, A., Martinez-Costa, C., Querol, A., Collado, M.C., Mira, A. (2017) Multiple approaches detect the presence of fungi in human breastmilk samples from healthy mothers. Sci Rep. 7:13016.
- Borneman, A.R., Desany, B.A., Riches, D., Affourtit, J.P., Forgan, A.H., Pretorius, I.S., *et al*. (2011) Whole-genome comparison reveals novel genetic elements that characterize the genome of industrial strains of Saccharomyces cerevisiae. PLoS Genet. 7:e1001287.
- Brand, S. (2009) Crohn's disease: Th1, Th17 or both? The change of a paradigm: new immunological and genetic insights implicate Th17 cells in the pathogenesis of Crohn's disease. Gut. 58:1152-1167.
- Cavalieri, D., McGovern, P.E., Hartl, D.L., Mortimer, R., Polsinelli, M. (2003) Evidence for *S. cerevisiae* fermentation in ancient wine. J. Mol. Evol.;57 Suppl 1:S226-32.
- Cavalieri, D., Di Paola, M., Rizzetto, L., Tocci, N., De Filippo, C., Lionetti, P., Ardizzoni, A., et al. (2018) Genomic and phenotypic variation in morphogenetic networks of two *Candida albicans* isolates subtends their different pathogenic potential. Front Immunol. 8:1997.
- Chiaro, T.R., Soto, R., Zac Stephens, W., Kubinak, J.L., Petersen, C., Gogokhia, L., *et al*. (2017) A member of the gut mycobiota modulates host purine metabolism exacerbating colitis in mice. Sci. Transl. Med. 9:eaaf9044.
- Dapporto, L., Stefanini, I., Rivero, D., Polsinelli, M., Capretti, P., De Marchi, P., *et al*. (2016) Social wasp intestines host the local phenotypic variability of *Saccharomyces cerevisiae* strains. Yeast. 33:277-87.
- Deutschbauer, A.M., Davis, R.W. (2005) Quantitative trait loci mapped to single-nucleotide resolution in yeast. Nat Genet. 37:1333-40.
- Diezmann, S., Dietrich, F.S. (2009) *Saccharomyces cerevisiae*: population divergence and resistance to oxidative stress in clinical, domesticated and wild isolates. Fay JC, editor. PLoS One 4:e5317.
- Dixon, P. (2003) VEGAN, a package of R functions for community ecology. J. Veg. Sci. 14:927–30.
- Dunn, B., Richter, C., Kvitek, D.J., Pugh, T., Sherlock, G. (2012) Analysis of the *Saccharomyces cerevisiae* pan-genome reveals a pool of copy number variants distributed in diverse yeast strains from differing industrial environments. Genome Res. 22:908–24.
- Fay, J.C., Benavides, J.A. (2005) Evidence for domesticated and wild populations of *Saccharomyces cerevisiae*. PLoS Genet. 1:e5.
- Hardison, S.E., Brown, GD. (2012) C-type lectin receptors orchestrate antifungal immunity. Nat. Immunol. 13:817–22.
- Ezov, T.K., Boger-Nadjar, E., Frenkel, Z., Katsperovski, I., Kemeny, S., Nevo, E., *et al*. (2006) Molecular-genetic biodiversity in a natural population of the yeast *Saccharomyces cerevisiae* from "Evolution Canyon": microsatellite polymorphism, ploidy and controversial sexual status. Genetics 174:1455–68.
- Findley, K., Oh, J., Yang, J., Conlan, S., Deming, C., Meyer, J.A., *et al*. (2013)Topographic diversity of fungal and bacterial communities in human skin. Nature 498:367-70.
- Gaffen, S.L., Hernández-Santos, N., Peterson, A.C. (2011) IL-17 signaling in host defense against *Candida albicans*.Immunol Res. 50:181-7.

- Gagliani, N., Huber, S. (2017) Basic aspects of T helper cell differentiation. Methods Mol Biol. 1514:19-30.
- Gerke, J., Lorenz, K., Cohen, B. (2009) Genetic interactions between transcription factors cause natural variation in yeast. Science 323:498–501.
- Gilbert, P.B., Novitsky, V.A., Montano, M.A., Essex, M. (2001) An efficient test for comparing sequence diversity between two populations.J Comput Biol. 8:123-39.
- Goddard, M.R., Anfang, N., Tang, R., Gardner, R.C., Jun, C. (2010) A distinct population of *Saccharomyces cerevisiae* in New Zealand: Evidence for local dispersal by insects and human-aided global dispersal in oak barrels. Environ. Microbiol. 12:63–73.
- Gozalbo, D., Maneu, V., Gil, M.L. (2014) Role of IFN-gamma in immune responses to *Candida albicans* infections. Front Biosci (Landmark Ed) 19:1279-90.
- Hallen-Adams, H.E., Suhr, M.J. (2017) Fungi in the healthy human gastrointestinal tract. Virulence 8:352–8.
- Hoffman, C.S., Winston, F. (1987) A ten-minute DNA preparation from yeast efficiently releases autonomous plasmids for transformation of *Escherichia coli*. Gene 57:267–72.
- Iliev, I.D., Funari, V.A., Taylor, K.D., Nguyen, Q., Reyes, C.N., Strom, S.P., *et al*. (2012) Interactions between commensal fungi and the C-type lectin receptor Dectin-1 Influence colitis. Science 336:1314–7.
- Knight, S., Klaere, S., Fedrizzi, B., Goddard, M.R. (2015) Regional microbial signatures positively correlate with differential wine phenotypes: evidence for a microbial aspect to terroir. Sci. Rep. 5:14233.
- Koh, A.Y. (2013) Gastrointestinal colonization of fungi. Curr. Fungal Infect. Rep. 7:144–51.
- Legras, J.L., Merdinoglu, D., Cornuet, J.M., Karst, F. (2007) Bread, beer and wine: *Saccharomyces cerevisiae* diversity reflects human history. Mol. Ecol. 16:2091–102.
- Lewis, L.E., Bain, J.M., Lowes, C., Gillespie, C., Rudkin, F.M., Gow, N.A.R., *et al.* (2012) Stage specific assessment of *Candida albicans* phagocytosis by macrophages identifies cell wall composition and morphogenesis as key determinants. PLoS Pathog. 8:e1002578.
- Li, H., Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–60.
- Li, Q., Wang, C., Tang, C., He, Q., Li, N., Li, J. (2013) Dysbiosis of gut fungal microbiota is associated with mucosal inflammation in Crohn's disease. J. Clin. Gastroenterol. 48:513-23.
- Liguori, G., Lamas, B., Richard, M.L., Brandi, G., da Costa, G., Hoffmann, T.W., *et al*. (2016) Fungal dysbiosis in mucosa-associated microbiota of Crohn's disease patients. J. Crohns. Colitis 10:296–305.
- Liti, G., Carter, D.M., Moses, A.M., Warringer, J., Parts, L., James, S.A., *et al.* (2009) Population genomics of domestic and wild yeasts. Nature 458:337–41.
- Liti, G. (2015) The fascinating and secret wild life of the budding yeast *S. cerevisiae*. Elife 4: e05835.
- Marakalala, M.J., Vautier, S., Potrykus, J., Walker, L.A., Shepardson, K.M., Hopke, A., et al. (2013) Differential adaptation of *Candida albicans* in vivo modulates immune recognition by dectin-1. PLoS Pathog. 9:e1003315.
- McCusker, J.H., Clemons, K.V., Stevens, D.A., Davis, R.W. (1994) *Saccharomyces cerevisiae* virulence phenotype as determined with CD-1 mice is associated with the ability to grow at 42 degrees C and form pseudohyphae. Infect. Immun. 62:5447–55.
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., *et al*. (2010) The genome analysis toolkit: a mapReduce framework for analyzing next-generation

DNA sequencing data. Genome Res. 20:1297–303.
- McKenzie, H., Main, J., Pennington, C.R., Parratt, D. (1990) Antibody to selected strains of *Saccharomyces cerevisiae* (baker's and brewer's yeast) and *Candida albicans* in Crohn's disease. Gut 31:536–8.
- Mortimer, R.K., Johnston, J.R. (1986) Genealogy of principal strains of the yeast genetic stock center. Genetics 113:35–43.
- Muñoz, P., Bouza, E., Cuenca-Estrella, M., Eiros, J.M., Pérez, M.J., Sánchez-Somolinos, M. *et al*. (2005) *Saccharomyces cerevisiae* fungemia: an emerging infectious disease. Clin Infect Dis. 40:1625-34.
- Nguyen, L.D.N., Viscogliosi, E., Delhaes, L. (2015) The lung mycobiome: an emerging field of the human respiratory microbiome. Front. Microbiol. 6:89.
- Novitsky, V.A., Montano, M.A., McLane, M.F., Renjifo, B., Vannberg, F., Foley, B.T., *et al*. (1999) Molecular cloning and phylogenetic analysis of human immunodeficiency virus type 1 subtype C: a set of 23 full-length clones from Botswana. J Virol. 73:4427-32.
- O'Meara, T.R., Veri, A.O., Ketela, T., Jiang, B., Roemer, T., Cowen, L.E. (2015) Global analysis of fungal morphology exposes mechanisms of host cell escape. Nat. Commun. 6:6741.
- Ott, S.J., Kühbacher, T., Musfeldt, M., Rosenstiel, P., Hellmig, S., Rehman, A., et al. (2008) Fungi and inflammatory bowel diseases: alterations of composition and diversity. Scand. J. Gastroenterol. 43:831–41.
- Patel, R.K., Jain, M. (2012) NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS One. 7:e30619.
- Perez-García, L., Diaz-Jimenez, D.F., Lopez-Esparza, A., Mora-Montes, H.M. (2011) Role of cell wall polysaccharides during recognition of *Candida albicans* by the innate immune system. J. Glycobiol. 1:1–7.
- Peter, J., De Chiara, M., Friedrich, A., Yue, J.X., Pflieger, D., Bergström, A., *et al*. (2018) J. Genome evolution across 1,011 Saccharomyces cerevisiae isolates. Nature 556:339-44.
- Price, M.N., Dehal, P.S., Arkin, A.P. (2010) FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490.
- Qin, Y., Zhang, L., Xu, Z., Zhang, J., Jiang, Y.Y., Cao, Y., Yan, T. (2016) Innate immune cell response upon *Candida albicans* infection. Virulence. 7:512-26.
- Quinton, J.F., Sendid, B., Reumaux, D., Duthilleul, P., Cortot, A., Grandbastien, B., *et al*. (1998) Anti-*Saccharomyces cerevisiae* mannan antibodies combined with antineutrophil cytoplasmic autoantibodies in inflammatory bowel disease: prevalence and diagnostic role. Gut 42:788–91.
- Raj, A., Stephens, M., Pritchard, J.K. (2014) fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. Genetics. 197:573–89.
- Ramazzotti, M., Berná, L., Stefanini, I., Cavalieri, D. (2012) A computational pipeline to discover highly phylogenetically informative genes in sequenced genomes: Application to *Saccharomyces cerevisiae* natural strains. Nucleic Acids Res. 40:3834-48.
- Reuter, M., Bell, G., Greig, D. (2007) Increased outbreeding in yeast in response to dispersal by an insect vector. Curr. Biol. 17:R81-3.
- Rizzetto, L., Kuka, M., De Filippo, C., Cambi, A., Netea, M.G., Beltrame, L., *et al*. (2010) Differential IL-17 production and mannan recognition contribute to fungal pathogenicity and commensalism. J. Immunol. 184:4258–68.
- Rizzetto, L., Giovannini, G., Bromley, M., Bowyer, P., Romani, L., Cavalieri, D. (2013) Strain dependent variation of immune responses to A. fumigatus: definition of pathogenic species.

PLoS One 8:e56651.
- Rizzetto, L., De Filippo, C., Cavalieri, D. (2014) Richness and diversity of mammalian fungal communities shape innate and adaptive immunity in health and disease. Eur. J. Immunol. 44:3166–81.
- Rizzetto, L., Ifrim, D.C, Morett,i S., Tocci, N., Cheng, .SC., Quintin, J. et al. (2016) Fungal chitin induces trained immunity in human monocytes during cross-talk of the host with *Saccharomyces cerevisiae*. J. Biol. Chem. 291:7961–72.
- Romani, L. (2011) Immunity to fungal infections. Nat Rev Immunol. 11:275-88.
- Romani, L., Zelante, T., Palmieri, M., Napolioni, V., Picciolini, M., Velardi, A., *et al*. (2015) The cross-talk between opportunistic fungi and the mammalian host via microbiota's metabolism. Semin. Immunopathol. 37:163–71.
- Roussey, J.A., Olszewski, M.A., Osterholzer, J.J. (2016) Immunoregulation in fungal diseases. Microorganisms. 4. pii: E47.
- Ruderfer, D.M., Pratt, S.C., Seidel, H.S., Kruglyak, L. (2006) Population genomic analysis of outcrossing and recombination in yeast. Nat. Genet. 38:1077–81.
- Sancho, D., Reis e Sousa, C. (2012) Signaling by myeloid C-type lectin receptors in immunity and homeostasis. Annu. Rev. Immunol. 30:491–529.
- Schacherer, J., Shapiro, J.A., Ruderfer, D.M., Kruglyak, L. (2009) Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. Nature 458:342–5.
- Schicho, R., Shaykhutdinov, R., Ngo, J., Nazyrova, A., Schneider, C., Panaccione, R., Kaplan, G.G., *et al.* (2012) Quantitative metabolomic profiling of serum, plasma, and urine by (1)H NMR spectroscopy discriminates between patients with inflammatory bowel disease and healthy individuals. J Proteome Res. 11:3344-57.
- Schulze, J., Sonnenborn, U. (2009) Yeasts in the gut: from commensals to infectious agents. Dtsch. Arztebl. Int. 106:837–42.
- Sebastiani, F., Barberio, C., Casalone, E., Cavalieri, D., Polsinelli, M. (2002) Crosses between *Saccharomyces cerevisiae* and *Saccharomyces bayanus* generate fertile hybrids. Res. Microbiol. 153:53–8.
- Sendid, B. (1998) Anti-*Saccharomyces cerevisiae* mannan antibodies in familial crohn's disease. Am. J. Gastroenterol. 93:1306–10.
- Sievers, F., Higgins, D.G. (2014) Clustal Omega, accurate alignment of very large numbers of sequences. Methods Mol. Biol. 1079:105–16.
- Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J.M., Birol, I. (2009) ABySS: a parallel assembler for short read sequence data. Genome Res. 19:1117–23.
- Smith, D., Metzgar, D., Wills, C., Fierer, J. (2002) Fatal *Saccharomyces cerevisiae* aortic graft infection. J Clin Microbiol. 40:2691-2.
- Smith, I.M., Christensen, J.E., Arneborg, N., Jespersen, L. (2014) Yeast modulation of human dendritic cell cytokine secretion: an in vitro study. PLoS One. 9:e96595.
- Sobel, J.D., Vazquez, J., Lynch, M., Meriwether, C., Zervos, M.J. (1993) Vaginitis due to *Saccharomyces cerevisiae*: epidemiology, clinical aspects, and therapy. Clin Infect Dis. 16:93-9.
- Sokol, H., Leducq, V., Aschard, H., Pham, H.P., Jegou, S., Landman, C., *et al*. (2017) Fungal microbiota dysbiosis in IBD. Gut. 66:1039-1048.
- Stanke, M., Diekhans, M., Baertsch, R., Haussler, D. (2008) Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 24:637–44.
- Stefanini, I., Dapporto, L., Legras, J.L., Calabretta, A., Di Paola, M., De Filippo, C., *et al*. (2012) Role of social wasps in *Saccharomyces cerevisiae* ecology and evolution. Proc.

Natl. Acad. Sci. 109:13398–403.
- Strati, F., Cavalieri, D., Albanese, D., De Felice, C., Donati, C., Hayek, J., *et al*. (2016a) Altered gut microbiota in Rett syndrome. Microbiome 4:41.
- Strati, F., Di Paola, M., Stefanini, I., Albanese, D., Rizzetto, L., Lionetti, P., *et al*. (2016b) Age and gender affect the composition of fungal population of the human gastrointestinal tract. Front Microbiol. 7:1227.
- Strope, P.K., Skelly, D.A., Kozmin, S.G., Mahadevan, G., Stone, E.A., Magwene, P.M., *et al*. (2015) The 100-genomes strains, an *S. cerevisiae* resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. Genome Res. 25:762–74.
- Swinne, D., Nolard, N., Van Rooij, P., Detandt, M. (2009) Bloodstream yeast infections: a 15-month survey.Epidemiol Infect. 137:1037-40.
- Takezaki, N., Nei, M. (1996) Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. Genetics. 144:389–399.
- Tamura, K. (2011) Ribosome evolution: emergence of peptide synthesis machinery. J. Biosci. 36:921–8.
- Tanaka, N., Awai, A., Bhuiyan, M.S., Fujita, K., Fukui, H., Takegawa, K. (1999) Cell surface galactosylation is essential for nonsexual flocculation in *Schizosaccharomyces pombe*. J. Bacteriol. 181:1356–9.
- Tawfik, O.W., Papasian, C.J., Dixon, A.Y., Potter, L.M. (1989) *Saccharomyces cerevisiae* pneumonia in a patient with acquired immune deficiency syndrome. J Clin Microbiol. 27:1689-91.
- Underhill, D.M., Iliev, I.D. (2014) The mycobiota: interactions between commensal fungi and the host immune system. Nat. Rev. Immunol. 14:405–16.
- van de Veerdonk, F.L., Gresnigt, M.S., Romani, L., Netea, M.G., Latgé, J.P. (2017) Aspergillus fumigatus morphology and dynamic host interactions. Nat Rev Microbiol. 15:661-674.
- Venables, W. N., Ripley, B. D. (2002) Modern applied statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.
- Williams, J.S., Mufti, G.J., Powell, S., Salisbury, J.R., Higgins, E.M. (2007) *Saccharomyces cerevisiae* emboli in an immunocompromised patient with relapsed acute myeloid leukaemia. Clin Exp Dermatol. 32:395-7.
- Xu, J., Boyd, C.M., Livingston, E., Meyer, W., Madden, J.F., Mitchell, T.G. (1999) Species and genotypic diversities and similarities of pathogenic yeasts colonizing women. J. Clin. Microbiol. 37:3835–43.
- Zhu, Y.O., Sherlock, G., Petrov, D.A. (2016) Whole genome analysis of 132 clinical *Saccharomyces cerevisiae* strains reveals extensive ploidy variation. Genes|Genomes| Genetics 6:2421–34.
- Zhu, Y.O., Sherlock, G., Petrov, D.A. (2017) Extremely rare polymorphisms in *Saccharomyces cerevisiae* allow inference of the mutational spectrum. Sunyaev SR, editor. PLOS Genet. 13:e1006455.

# Figure legends

**Figure 1: Human gut *Saccharomyces cerevisiae* strain typing.** *S. cerevisiae* strains isolated from faecal samples and identified by means of ITS1-5.8S-ITS2 sequencing were typed by characterizing 12 microsatellite loci. **a)** Neighbor-joining clustering based on the DC-chord distances calculated on the strains' microsatellite profiles. Shaded rectangles highlight the clusters identified through K-means analysis. Italic numbers indicate the cluster numbering reported in panel b; on the right, the main clusters including strains isolated from human faeces; grey numbers reported in correspondence of the nodes of these sub-clusters are the confidence values of the corresponding node according to bootstrap analysis; **b)** Comparison of chord distances, calculated from microsatellite profiles, among strains isolated from the same patient (B, D, E, G, H, P) or among strains clustering with the human faeces isolates. HG1, HG2, HG3, 8, and 13 refer to the clusters highlighted in panel a. Each grey circle represents the distance between two strains of the given group (either cluster or faecal donor); **c)** Results of statistical tests carried out to evaluate the differences in distances among strains grouped according to either donor or clustering. Mann-Whitney U test was carried out among inter-strain distances, the resulting p-values were corrected for multiple testing (false discovery rate). Differences were considered significant when fdr<0.05.

**Figure 2: Reconstruction of *S. cerevisiae* population phylogeny. a)** *S. cerevisiae* strain cluster based on the concatenated alignments of 1,715 informative CDSs (245,990 loci), obtained from whole genome sequences. Colour code of labels indicates the strains origin. Shaded areas indicate the inferred ancestry, as shown in panel b; **b)** The strains' most probable ancestry inferred by fastStructure analysis performed on whole-genome SNPs selected as having minor allele frequency>0.05 (28,473 loci). The proportions of the most probable inferred ancestors for each strain are indicated by the colours of vertical bars, as reported in the legend. The blue asterisk indicates the YA5-46A strain, isolated from human faeces and inferred to be a mosaic strain. Naming of clusters: sake - sake ancestor; WA - West Africa (as reported by Liti et al., 2009); HG - Human Gut; Lab - laboratory strains; WC - wild cluster; HB - Human Body; WE -Wine European.

**Figure 3: CNVs, gene loss prediction, discovery of new genes, and polymorphism frequencies of strains belonging to different clusters/ancestors.** Clustering of sequenced strains based on: a) gained genes, b) CNV's, c) defective genes, and d) polymorphism frequencies. The colour of the strain name indicates the strain's environmental origin, while the branch colour indicates the cluster in which the strain was found, as reported in Figure 1 (both colour scheme are shown in key). e) Survey of copy number variations (CNVs) and defective genes occurring along the chromosomes. Strains were grouped according to the clustering membership observed by means of ancestry analysis on genomic data. For each chromosome, two plots were generated: the upper reporting the average value of CNVs in CDSs, calculated in 5000 bp windows for strains in the corresponding group; the lower reporting the average number of defective CDSs calculated in 5000 bp windows in the corresponding group.

**Figure 4: Strains phenotyping.** Strains isolated from various environments were characterized in traits relevant for growth and survival in the gut environment: invasiveness, sporulation efficiency, resistance to several physiological temperatures and pH. Traits were quantified by Z-score, calculated as described in Experimental Procedures. Strain labels are coloured according to the isolation source (environment).

**Figure 5: Cell wall sugar components of *S. cerevisiae* strains.** Galactose, glucosamine, glucose and mannose amounts (%) detected in the cell wall of strains grown in presence of glucose (YPD) as carbon sources. The analysed strains are categorized considering (a) the different origins (environment), as indicated in figure, and (b) the inferred ancestors (Cluster). The graphs provide the p-values obtained by Mann-Whitney U test, corrected for multiple testing with the false discovery rate approach; *=fdr<0.05.

**Figure 6: Immunophenotyping of *S. cerevisiae* strains. a,b,c)** quantified amounts of IL-10, IL-17, and IFN-γ produced by PBMCs upon stimulation with *S. cerevisiae* cells; PBMCs isolated from

healthy donors were challenged with *S. cerevisiae* strains isolated from human faeces, grapes, other natural sources, and laboratory strains for 5 days and cytokines (IFN-γ, IL-10,and IL-17A) released in the supernatants were quantified. Bars in plots represent the mean ± standard deviation (error bars) of cytokine amounts from 6 healthy PBMCs donors; us: unstimulated cells. **d)** Correlation among genotyping, immunophenotyping and sporulation ability of HG *S. cerevisiae* strains. Correspondence analysis, carried out by using the *S. cerevisiae* strains as cases and T-polarizing cytokine levels produced by healthy human PBMCs in response to the strains (quantitative variables) and the categorized sporulation rates (high>25%, low<25%) as variables. Pie charts indicate the ancestor lineage proportion of each whole-genome sequenced strain (as reported in Cluster legend and coloured as in Figure 2), grey points represent strains with no genomic data. Ancestries were calculated for isolates or for meiotic segregants. If meiotic segregants deriving from the same parental strain showed different ancestries, a pie-chart was plotted for each segregant (e.g. YA5 derivatives), otherwise a single representative pie-chart was plotted (e.g. YH1 derivatives). Strains not characterized by genome sequencing but tested for immunophenotyping, were indicated with grey dots. In grey shadows, visual representation of strain grouping inferred using PAM analysis on PCA coordinates.

# Tables

## Table 1: Isolated human gut strains.

For CD (Crohn's Disease), UC (Ulcerative Colitis), nIBD (non-IBD Disease), and HC (Healthy Children) donor classes we reported the total number of donors and the number of *S. cerevisiae* strains isolated in the corresponding class. Strains marked with an asterisk (*) have been subjected to genome sequencing. Values in squared brackets indicate the number of sequenced meiotic segregants.

| Donor class | Donor ID | Strain Label |
|---|---|---|
| Crohn's Disease (CD)<br>n° subjects = 34,<br>having S. cerevisiae = 6 | A | YA5* [2*] |
| | B | YB7*, YB8*, YB9, YB10*, YB11, YB12, YB13 |
| | P | YP1, YP2, YP3, YP4 [1*] |
| | D | YD1*, YD2, YD3, YD4, YD5, YD6 |
| | E | YE1*, YE2*, YE3*, YE4, YE5 |
| | H | YH1* [3*], YH2, YH3, YH4 |
| Ulcerative Colitis (UC)<br>n° subjects = 27,<br>having *S. cerevisiae* = 1 | UC22 | YUC22* [1*] |
| non-IBD Disease (nIBD)<br>n° subjects = 1<br>having *S. cerevisiae* = 1 | G | YG9, YG12, YG13, YG14, YG15 |
| Healthy Children (HC)<br>n° subjects = 32,<br>having *S. cerevisiae* = 2 | N19 | YN19 |
| | EU13 | Y13EU* |

## Table 2: GO terms enrichment analyses.

Results of the GO Enrichment analysis performed on multiplied, lost, or defective CDS lists specific of the different clusters/groups investigated in this work. For each GO category, only terms enriched with an fdr < 0.05 are included.

| | | Gene Ontology | | |
|---|---|---|---|---|
| **Type** | **Cluster / group** | **Biological Process** | **Cellular Component** | **Molecular Function** |
| CNV higher than 1 | HG1 | none | integral component of plasma membrane [GO:0005887], intrinsic component of plasma membrane [GO:0031226], cell periphery [GO:0071944], plasma membrane part [GO:0044459],plasma membrane [GO:0005886] | inorganic molecular entity transmembrane transporter activity [GO:0015318], ion transmembrane transporter activity [GO:0015075], transmembrane transporter activity [GO:0022857] |
| | HG2 | allantoin metabolic process [GO:0000255], allantoin catabolic process [GO:0000256], cellular amide catabolic process [GO:0043605] | none | none |
| | HG3 | none | none | none |
| | intestinal | allantoin metabolic process [GO:0000255], allantoin catabolic process [GO:0000256] | none | none |
| | non-intestinal | none | cell periphery [GO:0071944], integral component of plasma membrane [GO:0005887], intrinsic component of plasma membrane [GO:0031226], plasma membrane [GO:0005886], plasma membrane part [GO:0044459] | none |
| CNV lower than -1 (lost) | HG1 | none | none | none |
| | HG2 | none | none | none |
| | HG3 | none | none | none |
| | intestinal | none | cell wall [GO:0005618], external encapsulating structure [GO:0030312], fungal-type cell wall [GO:0009277], cell periphery [GO:0071944], vacuole [GO:0005773] | structural constituent of cell wall [GO:0005199] |
| | non-intestinal | none | none | none |
| defective genes | HG1 | none | cell periphery [GO:0071944] | none |
| | HG2 | none | none | none |
| | HG3 | none | none | none |
| | intestinal | iron ion homeostasis [GO:0055072], metal ion homeostasis [GO:0055065], cellular iron ion homeostasis [GO:0006879] | none | none |
| | non-intestinal | none | none | none |